Towards Automated RISC-V Microarchitecture Design with Reinforcement Learning

Chen Bai[1], **Jianwang Zhai**[2†], **Yuzhe Ma**[3], **Bei Yu**[1†], **Martin D.F. Wong**[4]
[1]The Chinese University of Hong Kong
[2]Beijing University of Posts and Telecommunications
[3]The Hong Kong University of Science and Technology (Guangzhou)
[4]Hong Kong Baptist University
{cbai, byu}@cse.cuhk.edu.hk, zhaijw@bupt.edu.cn, yuzhema@hkust-gz.edu.cn, mdfwong@hkbu.edu.hk

Abstract

Microarchitecture determines the implementation of a microprocessor. Designing a microarchitecture to achieve better performance, power, and area (PPA) trade-off has been increasingly difficult. Previous data-driven methodologies hold inappropriate assumptions and lack more tightly coupling with expert knowledge. This paper proposes a novel reinforcement learning-based (RL) solution that addresses these limitations. With the integration of microarchitecture scaling graph, PPA preference space embedding, and proposed lightweight environment in RL, experiments using commercial electronic design automation (EDA) tools show that our method achieves an average PPA trade-off improvement of 16.03% than previous state-of-the-art approaches with $4.07\times$ higher efficiency. The solution qualities outperform human implementations by at most $2.03\times$ in the PPA trade-off.

Introduction

The instruction set architecture (ISA) is the interface between software and hardware. RISC-V, an open standard ISA, has garnered significant attention from both academia and industry nowadays. The microarchitecture determines how a particular microprocessor is implemented given an ISA. It sets the cornerstone for a microprocessor's overarching design points: performance, power, and area (PPA).

Nevertheless, it is challenging to design a microarchitecture efficiently to achieve pre-determined PPA design goals for target workloads (computation-bound or memory-bound programs) with manual efforts. Computer architects often rely on design space exploration (DSE) to find appropriate solutions. Those solutions can maximize performance and minimize power and area for target workloads. DSE is an iterative, trial-and-error, and non-trivial procedure due to two factors. First, the design space is enormous and complicated. It comes from the high complexity of a microarchitecture, such as shown in Figure 1, which includes different *components* responsible for implementing specific functions (Chen et al. 2020; Grayson et al. 2020). Second, evaluating the PPA values of a single microarchitecture design requires an extremely high runtime. Thus, it is a fond dream for architects to iterate the design space and retrieve optimal solutions.

† Corresponding Authors

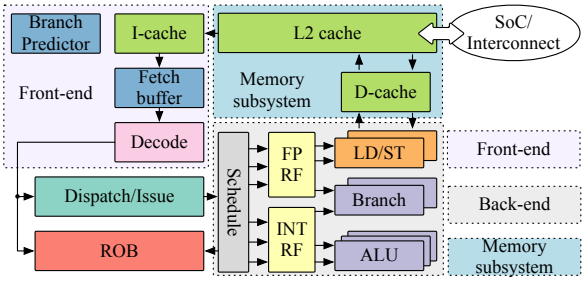

Figure 1: Instructions fetched from I-cache are sent to functional units (e.g., ALU, LD/ST, etc.) for execution. Register files (RF) save temporary data. Reorder buffer (ROB) achieves precise interrupts (Smith and Pleszkun 1988). Components are highlighted with diverse colors, and the same color denotes a similar function.

Previous methodologies have been proposed. In industry, architects' expert knowledge is a heuristic to guide the DSE. However, it is a concern that personal bias can lead to suboptimal solutions. In academia, both analytical and data-driven methods have been proposed. The analytical methods conduct interpretable equations to describe relations between microarchitectures and PPA values for various workloads. Karkhanis and Smith (2007) adopted interval analysis to construct such equations. However, the analytical model requires much expert knowledge, and is unscalable for newly-emerged microprocessors. Data-driven methods are utilized accordingly when we lack accesses to experts. The microarchitecture is viewed as a black box. Chen et al. (2014) employed a ranking model. Li et al. (2016) applied statistical sampling and AdaBoost learning. Bai et al. (2021; 2023b) proposed a Bayesian optimization-based framework. Such data-driven methods generally outperform analytical methods owing to many advanced machine-learning techniques (Li et al. 2022; Yu et al. 2023). However, they are not free of criticism. Blindly exploring microarchitectures (purely driven by the algorithm rather than tightly coupled with expertise) can be naive since architects already know the characteristics of most designs (Bai et al. 2023a).

In this paper, we follow the approach of previous data-driven methods but with key distinctions: our method removes prior unrealistic assumptions, and our solution is deeply integrated with expert knowledge. Previous data-

driven methods often assume a positive correlation between the PPA difference and feature embeddings of microarchitectures. On the contrary, the assumption does not hold in general. Our RL solution is free of such assumptions. Moreover, using the *microarchitecture scaling graph*, we tightly embed the expert knowledge to formulate the Markov decision process (MDP). The scaling graph encodes the sequential decision precedences of the microarchitecture components. Accordingly, we propose a multi-objective RL framework based on the MDP. The framework enables the automated RISC-V microarchitecture design with a single agent for different PPA design preferences. It is worth noting that our solution focuses on RISC-V to promote chip agile design methodology (Lee et al. 2016). Our main contributions are as follows:

1) We propose an MDP model with the microarchitecture scaling graph, embracing architects' expertise and providing strong prior knowledge for our agent.

2) We embed the PPA design preferences into RL and re-formulate the multi-objective optimization to a unified dynamic-weighted reward signal. It is helpful since this feature allows agents to explore microarchitectures for different PPA design preferences online.

3) We propose a lightweight environment to accelerate the learning process. With calibrated PPA models, we accelerate the learning process by over $100\times$ times compared to using EDA tools only (Zhai et al. 2021).

4) Our experiments use representative RISC-V microprocessors and evaluate with commercial EDA tools at 7-nm technology. Results show that our method can achieve an average of 16.03% PPA trade-off improvement over prior state-of-the-art approaches with $4.07\times$ higher efficiency. And the solution qualities outperform human implementations by at most $2.03\times$ in the PPA trade-off.

Preliminary

RISC-V Microprocessors. Unlike other ISAs (ARM, x86, etc.), RISC-V is free for commercial usage. The free license drives the appearance of many RISC-V microprocessors, some of which are representatives. Rocket (Asanović et al. 2016) is a six-stage pipeline in-order microprocessor. SonicBOOM (Zhao et al. 2020) is a ten-stage pipeline out-of-order design. XiangShan (Xu et al. 2022) features advanced microarchitecture optimizations. Xuantie-910 (Chen et al. 2020) is an open-source implementation from the industry.

Microarchitecture PPA Modeling. Computer architects use various tools to evaluate PPA values of microarchitecture designs. When the register-transfer-level design (RTL) [1] is available, EDA tools are necessary to report PPA values. The PPA evaluation flow with EDA tools includes steps like logic synthesis, placement and routing, netlist simulation, etc. When the RTL implementation is unavailable, pre-RTL simulation infrastructures like microprocessor performance simulators are used to report first-hand PPA values. Compared to EDA tools, pre-RTL simulation infrastructures are less accurate. In this paper, we propose a lightweight RL environment to couple the pre-RTL simulation infrastructures with EDA tools, i.e., improve the modeling accuracy of pre-RTL simulation infrastructures without sacrificing efficiency. Specifically, we leverage GEM5 (Binkert et al. 2011), a performance simulator, and McPAT (Li et al. 2009) as our fundamental PPA modeling tools.

[1] Register-transfer level design (RTL) is a description of hardware implementations using programming languages such as Verilog and VHDL.

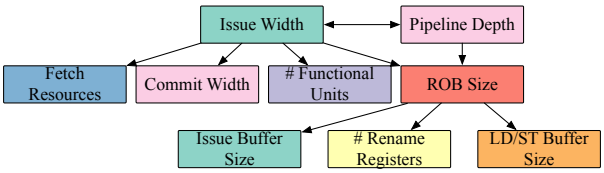

Figure 2: An overview of the microarchitecture scaling graph. Colors are matched with Figure 1.

Microarchitecture Scaling Graph. Since the microarchitecture scaling graph first appeared (Eyerman et al. 2009), computer architects relied on it to study mechanistic microexecutions (Carlson, Heirman, and Eeckhout 2011; Carlson et al. 2014).

The microarchitecture scaling graph is directed, elucidating the scaling precedence constraints between components, as shown in Figure 2. Nodes are components, and directed edges are scaling precedences. According to Figure 2, an interplay exists between the pipeline width and issue width. The pipeline width and issue width determine the ROB size, while the ROB size decides the issue buffer size, load/store (LD/ST) buffer size, etc. The scaling graph is derived from extensive simulations and architects' discussions. The relations unfolded by the scaling graph are general for mainstream microarchitectures due to a widely applied typical von Neumann architecture.

Problem Formulation. Given the microarchitecture design space, the problem is to find the solution to Equation (1) within a limited time budget.

$$\underset{\boldsymbol{s}\in\mathbb{D}^n}{max}[\text{Perf}(\boldsymbol{s}), -\text{Power}(\boldsymbol{s}), -\text{Area}(\boldsymbol{s})], \tag{1}$$

where $\mathbb{D}$ is an n-dimensional microarchitecture design space, and $\boldsymbol{s}$ is a vector to parameterize a design (feature embedding of a microarchitecture).

Methodology

Overview

We propose an RL solution framework with customized MDP (S, A, P, R) formulation, as shown in Figure 3. The state space S is the design space. The action space A is the candidate set of components' types or corresponding hardware resources listed in Table 1. The components' types refer to a type of branch predictor, cache replacement policy, etc., and hardware resources specify the queue, buffer, or stack sizes. P is the state transition. The reward space R involves all the vectorized PPA values. In the training, the agent learns to generate appropriate *partial* components stepwise given sampled PPA preference vectors to formulate the complete microarchitecture. In the DSE, the trained agent produces solutions w.r.t. a fixed PPA preference specified by architects.

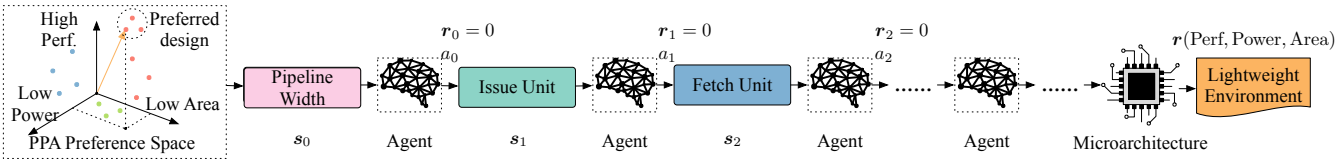

Figure 3: Overview of our RL framework. $\boldsymbol{s}$ denotes a state, a is an action, and $\boldsymbol{r}$ represents an immediate vectorized reward.

The PPA preference space is incorporated into our framework since the microarchitecture design is faced with diverse workloads. High-performance computing scenarios emphasize performance more, while embedded applications push pressure on high power efficiency and area efficiency. Different PPA preference vectors denote architects' design goals, and our single agent benefits from the PPA preference-aware DSE with the proposed framework.

Combine RL w. Microarchitecture Scaling Graph

We combine RL with the microarchitecture scaling graph via a practical episode design. In each step of an episode, the agent produces partial components. The episode ends until a complete microarchitecture is formulated.

The state of a microarchitecture is encoded as a vector with each element denoting a selection of a particular component parameter. Elements are masked for undefined components and masks are removed progressively as more components are determined by advancing the step. The action space is correlated with the step since different components relate to distinct action candidates. Each step determines one component. Once the complete microarchitecture is generated, we adopt the lightweight environment to evaluate the reward $\boldsymbol{r}$(Perf, Power, Area). Otherwise, the reward is zero. The precedence of decision-making among components follows the scaling graph, as the graph unveils cause-and-effect relations between different components. Another noteworthy point is that the action space is changed in each step, leading to the output misalignment for a single agent. We apply a typical engineering trick: the normalization of action probability to deal with it (Schrittwieser et al. 2020).

The rationale for the episode design is that we place strong prior knowledge for the agent. The prior knowledge is derived from the scaling graph, revealing the components' decision priority. For example, the pipeline width determines the maximal instructions fetched simultaneously. And the structure of the issue unit is then decided based on the width. Because the issue unit adjusts instruction issue rates based on how many instructions are fetched by the front-end and buffers those instructions for the back-end (see Figure 1). Hence, our episode design explicitly provides such domain knowledge for the agent. As shown in Figure 3, once a PPA preference and the pipeline width are specified, the agent first determines the appropriate issue queue sizes, with the following determinations involving fetch queue size, type of a branch predictor, etc., sequentially. Although relations between some components are not uncovered from the scaling graph (e.g., the issue buffer size and the number of physical registers indicated in Figure 2), we determine their structures in a fixed order within an episode.

Dynamic-weighted Reward

Since a single agent cannot handle multiple objectives simultaneously, a weighted summation is applied in the reward computation, as listed in Equation (2).

$$r = \boldsymbol{r}(\text{Perf}, \text{Power}, \text{Area}) \cdot (\alpha, \beta, \gamma)^\top \tag{2}$$

where α, β, and γ are weights controlling the PPA trade-off. We align the reward optimization with our objectives via normalizing Perf, Power, and Area, i.e., maximizing the reward equals maximizing the performance, and minimizing the power and area.

However, weights can be changed as architects' PPA design goals vary. A transparent limitation is that the agent needs to be retrained once the weights are changed. Accordingly, an online adaptation for changed weights is necessary. That is, the single agent can handle the changing coefficients α, β, and γ without learning from scratch. It motivates us to embed the PPA preference space into the framework.

Embed Preference Space into RL

The PPA preference space Φ is the set of preference vectors $\boldsymbol{\phi} = (\alpha, \beta, \gamma)$, which balance the PPA values in various degrees and satisfy the simplex constraints, i.e., $\forall i,\ \boldsymbol{\phi}_i \geq 0,\ \sum_i \boldsymbol{\phi}_i = 1$. We embed Φ into RL, making the agent learn the *convex coverage set* (CCS) w.r.t. Equation (2) (Roijers, Whiteson, and Oliehoek 2015; Roijers and Whiteson 2017). Hence, a single agent can maximize r without retraining or fine-tuning in the DSE when $\boldsymbol{\phi}$ is changed.

CCS is the convex subset of the Pareto frontier, as formulated in Equation (3).

$$\begin{aligned} \text{CCS} = \{\boldsymbol{r} \in \mathcal{PF}(R) \mid \\ \exists \boldsymbol{\phi} \in \Phi, \boldsymbol{r}\boldsymbol{\phi}^\top \geq \boldsymbol{r}'\boldsymbol{\phi}^\top, \forall \boldsymbol{r}' \in \mathcal{PF}(R)\}, \end{aligned} \tag{3}$$

where $\mathcal{PF}(R)$ is the *Pareto frontier* of R, and $\mathcal{PF}(R) = \{\boldsymbol{r} \mid \nexists \boldsymbol{r}' \succeq \boldsymbol{r}, \forall \boldsymbol{r}, \forall \boldsymbol{r}' \in R\}$ [2]. Pareto frontier is a set of microarchitectures whose PPA values represent the best trade-off. Equation (3) indicates that optimal solutions can attain the maximal r (see Equation (2)) for a specific $\boldsymbol{\phi}$.

To facilitate the agent's learning of the CCS during training, a generalized Bellman optimality equality is applied (Yang, Sun, and Narasimhan 2019). The generalized equality is an extension from the single objective Bellman optimality. The main idea is to optimize the policy towards maximal r given a particular $\boldsymbol{\phi}$, as shown in Equation (4).

$$\begin{aligned} \boldsymbol{Q}(\boldsymbol{s}, a, \boldsymbol{\phi}) &= \boldsymbol{r}(\boldsymbol{s}, a) + \zeta \mathbb{E}_{\boldsymbol{s}' \sim \mathcal{P}(\cdot|\boldsymbol{s},a)} \mathcal{T}(\boldsymbol{Q}(\boldsymbol{s}', a, \boldsymbol{\phi})), \\ \mathcal{T}(\boldsymbol{Q}(\boldsymbol{s}', a, \boldsymbol{\phi})) &= \underset{\boldsymbol{Q}}{arg} \underset{a' \in A, \boldsymbol{\phi}' \in \Phi}{max} \boldsymbol{Q}(\boldsymbol{s}', a', \boldsymbol{\phi}')\boldsymbol{\phi}^\top, \end{aligned} \tag{4}$$

[2] $\boldsymbol{r}' \succeq \boldsymbol{r}$ denotes that each element of $\boldsymbol{r}'$ is better than $\boldsymbol{r}$.

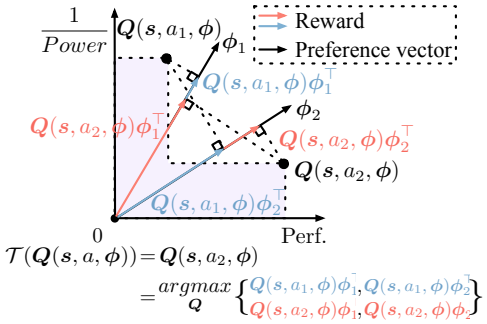

Figure 4: Optimization procedure with Equation (4).

where ζ is a discount factor. $\boldsymbol{Q}(\boldsymbol{s},a,\boldsymbol{\phi})$ is the state-action vector when $\boldsymbol{s}$ is the state, a is the action, and $\boldsymbol{\phi}$ is the preference vector. $\mathcal{T}(\boldsymbol{Q}(\boldsymbol{s},a,\boldsymbol{\phi}))$ is $\boldsymbol{Q}$, which attains the maximal r via traversing the action space A, and sampled $\boldsymbol{\phi}'$ [3].

Figure 4 details an example optimization with Equation (4) in the performance-power space, given $\boldsymbol{\phi}$. Before applying Equation (4), we sample multiple different $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$, holding an insight that the agent can learn to generate other policies according to varied preferences. At state $\boldsymbol{s}$, the agent is faced with actions a_1 and a_2. Under $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$, four rewards are highlighted with blue and red colors. Ultimately, the policy is optimized with $\boldsymbol{Q}(\boldsymbol{s},a_2,\boldsymbol{\phi})$ since it achieves the maximal reward among all rewards.

Our RL framework adopts the asynchronous advantage actor-critic (A3C) (Mnih, Badia et al. 2016) in favor of high training efficiency over PPO (Schulman et al. 2017) or SAC (Haarnoja et al. 2018). The actor is a policy network used to generate an action. The critic evaluates the complete state to determine whether the optimization becomes better or worse than expected. The gradients of the actor $\boldsymbol{\theta}_a$ is listed in Equation (5).

$$\begin{aligned}\bigtriangledown\boldsymbol{\theta}_a =&\kappa \bigtriangledown_{\boldsymbol{\theta}_a} H(\pi(\boldsymbol{s}_t;\boldsymbol{\theta}_a))+\\ &\mathbb{E}_{\xi\sim\pi}[\sum_{t=0}^{\infty}\bigtriangledown_{\boldsymbol{\theta}_a}\log\pi_{\boldsymbol{\theta}_a}(a_t\mid\boldsymbol{s}_t)A(\boldsymbol{s}_t,a_t,\boldsymbol{\phi}')\boldsymbol{\phi}^\top],\end{aligned} \tag{5}$$

where $A(\boldsymbol{s}_t,a_t,\boldsymbol{\phi}')=Q(\boldsymbol{s}_t,a_t,\boldsymbol{\phi}')-V(\boldsymbol{s}_t,\boldsymbol{\phi}')$ is the advantage function featuring relatively low variance, ξ is a trajectory following the policy π, and $\boldsymbol{\theta}_a$ denotes parameters of the actor. The entropy of the policy π is incorporated in optimizing the actor ($H(\pi(\boldsymbol{s}_t;\boldsymbol{\theta}_a))$. It can prevent the agent from always selecting the currently found best action. A coefficient κ controls the strength of entropy regularization. For the critic, Equation (6) gives the loss function with an L2 normalization applied between two state-action vectors.

$$\begin{aligned}L_c =&\rho\|(\boldsymbol{Q}^*-\boldsymbol{Q}(\boldsymbol{s},a,\boldsymbol{\phi}';\boldsymbol{\theta}_c))\boldsymbol{\phi}^\top\|_2^2+\\ &(1-\rho)\|\boldsymbol{Q}^*-\boldsymbol{Q}(\boldsymbol{s},a,\boldsymbol{\phi}';\boldsymbol{\theta}_c)\|_2^2,\end{aligned} \tag{6}$$

where ρ is a coefficient to balance these two terms, $\boldsymbol{\theta}_c$ denotes parameters of the critic, and $\boldsymbol{Q}^*$ is obtained from Equation (4). The first term in Equation (6) enforces optimizing the critic network w.r.t. the maximal reward shown in Equation (2). The *n-step* TD errors (Peng and Williams 1994) is leveraged. However, Equation (5) requires many transition samples to give a relatively accurate gradients approximation for a steady and stable improvement. We employ the generalized advantage estimator (GAE) to handle it (Schulman et al. 2016), as listed in Equation (7).

$$\boldsymbol{r}_t=\sum_{n=0}^{N}(\lambda\zeta)^{N-n}(\boldsymbol{r}_{t+k}+\zeta V_{t+1+k}(\boldsymbol{s}_t,\boldsymbol{\phi}')-V_{t+k}(\boldsymbol{s}_t,\boldsymbol{\phi}')), \tag{7}$$

where λ is a coefficient controlling the strength of the exponential-weighted average.

Conditioned Actor-Critic Network

The input of our actor and critic networks is the concatenation of state and corresponding preference vectors. The preference vectors serve as conditional inputs to the actor and critic networks. Both networks are multilayer perceptrons with leaky ReLU as the activation function. The intuition of the concatenation is to support the online adaptation of changed preferences for agents. Hence, many policies are optimized on-the-fly (Abels et al. 2019).

Accelerate Learning via Lightweight Environment

Training the agent with pre-RTL simulation infrastructures as the RL environment is inaccurate while using EDA tools in the loop is inefficient. So, we propose a "lightweight" environment to combine the merits of both modeling flows, which the "lightweight" refers to that our environment can achieve a speed-up of $100\times\ \sim\ 110\times$ in PPA estimation compared to using EDA tools in the training loop.

The lightweight environment is based on the calibration, which is set up before the RL training (Zhai et al. 2021). We leverage the EDA flow in the calibration as a PPA ground truths generation flow. And we adopt the pre-RTL simulation infrastructures as feature extraction flow. The extracted features encode knowledge to microarchitectures and workloads, e.g., queue, buffer, or stacks' number of reads and writes, number of load and store instructions, etc. Supervised learning is then applied to train PPA black-box models such as XGBoost (Chen and Guestrin 2016) separately, with the loss function defined in Equation (8).

$$L=|f(\boldsymbol{s},\boldsymbol{e},\boldsymbol{p})-y_{\text{gt}}|_2^2, \tag{8}$$

where f is a black-box model. $\boldsymbol{s}$, $\boldsymbol{e}$, and $\boldsymbol{p}$ are inputs of the model, denoting state, cycle accurate simulation statistics, and other PPA-related features (e.g., leakage and subthreshold power, etc.), respectively. y_{gt} is the ground truth. Figure 5 provides an overview of the calibration flow. In the RL training and DSE, a microarchitecture is initially evaluated using pre-RTL simulation infrastructures and is calibrated with trained PPA black-box models. Furthermore, we duplicate the environment into multiple instances, permitting higher training parallelism.

[3]The size of A at each step is small, allowing us to traverse efficiently. However, since Φ is an uncountable set, we sample multiple $\boldsymbol{\phi}$ in the training and compute $\mathcal{T}(\boldsymbol{Q}(\boldsymbol{s}',a,\boldsymbol{\phi}))$ based on these samples otherwise.

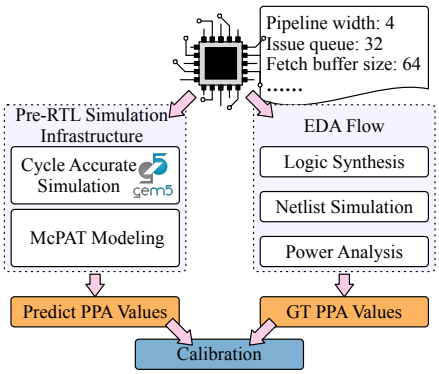

Figure 5: Overview of the PPA calibration.

Why RL?

Previous data-driven methods apply statistical analysis (Li et al. 2016), Gaussian process (Bai et al. 2021), etc. However, a limitation can be observed. *Most previous methods attribute the degree of PPA difference to the distance between feature embeddings of microarchitectures.* For example, the Gaussian process assumes the existence of such relations (Bai et al. 2021; Williams and Rasmussen 2006). On the contrary, we find the relation does not hold generally, and demonstrate it with an anti-example shown in Figure 6. M1 is the baseline microarchitecture. M2 changes the branch predictor (Seznec and Michaud 2006), M3 reduces the decode width, and M4 decreases branch speculation tags. t-SNE (Van der Maaten and Hinton 2008) is utilized to visualize the embedding distances to M1. Notwithstanding that M2 and M3 have the same distance to M1, they incur different PPA value gaps to M1. M3 has 8.54%, 3.00%, and 5.09% smaller PPA values than M1. M2 demonstrates a more substantial difference, i.e., 13.09%, 23.75%, and 14.48% lower PPA values than M1. The embedding distance between M1 and M2 is closer than that between M1 and M4. However, compared with M2, M4's PPA values are even closer to M1, i.e., M1 outperforms IPC by 0.36% [4], dissipating 3.67% more power and 1.39% larger area than M4.

Our RL solution can remove unrealistic assumptions. Thus, it alleviates the limitations of prior data-driven methods. However, RL might be one of many remedies while our MDP formulation can capture the structure of the problem.

Experiments

RISC-V Microarchitecture Design Space

We evaluate the proposed RL framework with representative in-order and out-of-order RISC-V microprocessors, Rocket (Asanović et al. 2016) and different scales of SonicBOOM (Zhao et al. 2020) (categorized by "pipelineWidth"), as shown in Table 1. We include cache structures, branch predictors, functional units, load/store units, issue units, etc., in the design space. The Rocket and SonicBOOM design space size is 5.18×10^6 and 1.02×10^{16}, respectively.

[4]Instruction per cycle (IPC) is a performance metric.

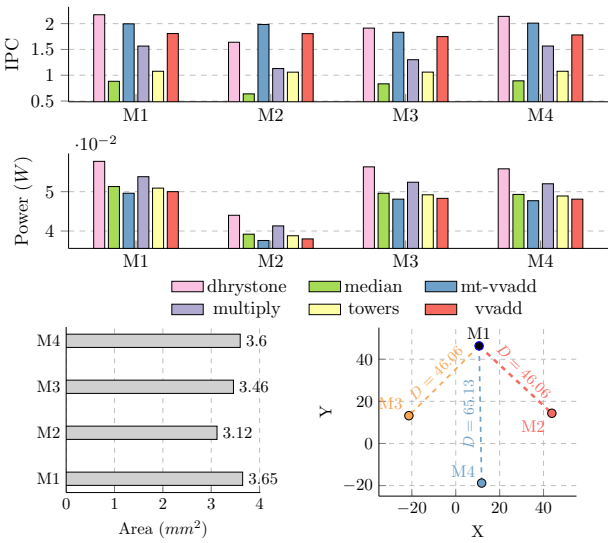

Figure 6: Four SonicBOOM microarchitectures' PPA values of six benchmarks reported from EDA tools, and the visualization of embeddings distances.

Experimental Settings & Baselines

All experiments are conducted on 80 Quad Intel(R) Xeon(R) CPU E7-4820 V3 cores with a 1 TB main memory. The PPA values reported in the main results are from commercial EDA tools. Specifically, the performance, power, and area values are obtained from Synopsys VCS M-2017.03, Synopsys PrimeTime PX R-2020.09-SP1, and Cadence Genus 18.12-e012_1 with 7-nm technology (Clark et al. 2016). Due to the page limit, we mainly show the results of SonicBOOM for some experiments, specifically for the Large SonicBOOM. Code is publicly available at https://github.com/baichen318/rl-explorer.

The coefficient κ in Equation (5) is set as 1, ρ in Equation (6) is 0.5, λ in Equation (7) is 0.95 and the discount factor ζ in Equation (7) is 0.99. Adam optimizer is used, and the initial learning rate is 0.001.

We compare our method with current state-of-the-arts, i.e., Bayesian optimization-based (Bai et al. 2021) (ICCAD'21), Adaboost-based (Li et al. 2016) (DAC'16), ranking-based (Chen et al. 2014) (ISCA'14), and human efforts (Asanović et al. 2016; Zhao et al. 2020). The baselines are implemented according to the original papers. We use `towers`, `vvadd`, `spmv` from official RISC-V tests as workloads in the DSE. Results on more workloads are also elucidated. To compare the efficiency of algorithms fairly, all baselines adopt the lightweight environment, but searched solutions are re-evaluated with EDA tools.

Accuracy of Lightweight PPA Models

We use the Kendall τ and the mean absolute percentage error (MAPE) to measure the accuracy of lightweight PPA models. The higher the Kendall τ and the lower the MAPE, the more accurate the lightweight PPA models are.

Figure 7 lists the accuracy of lightweight PPA XGBoost models and the ratio of microarchitectures in the training data set leveraged in the calibration flow. In the first row, the

Design	Component	Parameters	Candidate
Rocket	Branch predictor	RAS	$0:12:3^{+}$
		BTB.nEntries	$0:56:14$
		BHT.nEntries	$0:1024:256$
	I-cache	nWays	$1,2,4$
		nTLBWays	$4:32:4$
	Functional unit	FPU	$1,2$
		mulDiv	$1,2,3$
		VM	$1,2$
	D-cache	nSets	$32,64$
		nWays	$1,2,4$
		nTLBWays	$4:32:4$
		nMSHRs	$1,2,3$
Small Medium Large Mega Giga SonicBOOM	Branch predictor	Type	$1,2,3$
		maxBrCount	$4:22:2$
	IFU	fetchBufferEntries	$6:46:2$
		fetchWidth	$4,8$
		ftq.nEntries	$12:64:4$
	pipelineWidth		$1:5:1$
	ROB		$24:160:4$
	PRF	intPhysRegisters	$40:176:8$
		numFpPRF	$34:132:6$
	ISU	fpPhysRegisters	$1:5:1$
		numEntries	$6:52:2$
		dispatchWidth	$1:5:1$
	LSU	LDQ	$6:32:2$
		STQ	$6:36:2$
	I-cache	nWays	$4,8$
		nSets	$32,64$
	D-cache	nWays	$4,8$
		nSets	$64,128$
		nMSHRs	$2:10:2$

* The values are start number:end number:stride, e.g., $0:12:3$ denotes the entries of RAS can be 0, 3, 6, etc., until 12.

Table 1: RISC-V Microarchitecture Design Space

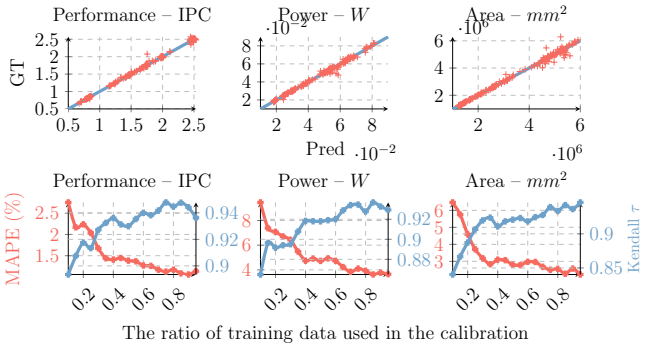

Figure 7: The accuracy of lightweight PPA models, and MAPE and Kendall τ curves w.r.t. the calibration data size.

blue line "GT = Pred" visualizes the error when PPA models are trained using the entire training data set. The Kendall τ are 0.93, 0.95, and 0.94 for PPA models, respectively, indicating acceptable accuracy when using these models in the RL framework. The second row demonstrates how much data the calibration flow needs to receive acceptably accurate PPA models. By leveraging around $800 \sim 900$ SonicBOOM microarchitecture designs, the Kendall τ for PPA modeling results can achieve higher than 0.92.

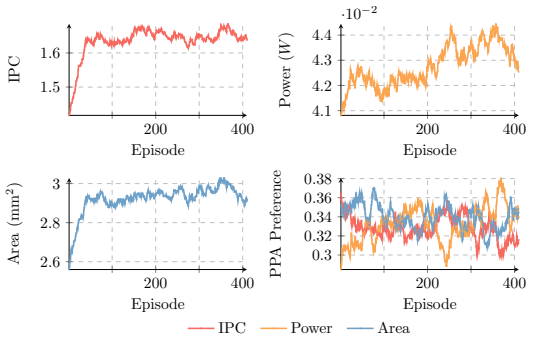

Figure 8: RL training status of Large SonicBOOM.

RL Training

Figure 8 displays the RL training metric curves for Large SonicBOOM, which include PPA values, and specific values of PPA preference vectors. Different PPA preference vectors are sampled throughout the training, resulting in perturbed PPA values received in each episode. IPC curves are increased gradually and flattened eventually as trained with more episodes. Power and area values are changing in response to divergent PPA preference vectors.

Comparison w. Human Efforts & Prior Arts

Three metrics: Perf/Power, Perf/Area, and (Perf $\times$ Perf)/(Power $\times$ Area) are used in the experiments. These metrics measure how much performance per watt, performance per area, and PPA trade-off a microprocessor can attain. The higher the values, the better the power/area efficiency of the microprocessor. In the DSE, we use predetermined PPA preference vectors for different scales of SonicBOOM. The preference vectors are $(1/12, 1/12, 10/12)$, $(1/7, 1/7, 5/7)$, $(1/3, 1/3, 1/3)$, $(5/7, 1/7, 1/7)$, and $(10/12, 1/12, 1/12)$ for Small, Medium, Large, Mega, and Giga SonicBOOM, respectively. We use $(1/3, 1/3, 1/3)$ for Rocket. These preference vectors are used in the RL to identify optimal designs and to calculate scalar rewards for solutions obtained through the baseline algorithms. We facilitate fair comparisons between different methodologies by comparing the solutions that yield the maximal reward among our method and the baseline algorithms in the abovementioned three metrics. The rationale behind setting such preference vectors is to emphasize specific design priorities based on the scale of the microprocessors. We prioritize higher power and area efficiency for small microprocessors, as reflected in the preference vectors. On the other hand, for larger microprocessors, we emphasize higher performance. For the middle scale of SonicBOOM, i.e., Large SonicBOOM, we aim for a higher degree of balance among the PPA values.

Table 2 lists the results. The relative runtime for exploration is also reported. Explored Rocket and nearly all scales of SonicBOOM by our method are better than prior works and human efforts. In summary, our solutions achieve an average of 24.64%, 17.13%, and 6.33% than ICCAD'21, DAC'16, and ISCA'14 in PPA trade-off, respectively. Moreover, the RL solutions outperform human efforts up to $2.03\times$ better in (Perf $\times$ Perf)/(Power $\times$ Area). And our method can

Design	Method	Performance IPC	Power mW	Area mm^2	Perf / Power		Perf / Area		(Perf × Perf) / (Power × Area)		Runtime
					Val.	Ratio	Val.	Ratio	Val.	Ratio	
Rocket	Human Efforts	0.734	2.700	0.908	0.272	– [1]	0.808	–	0.220	–	–
	ISCA'14	0.816	2.300	0.794	**0.355**	**1.305**×	1.027	1.271×	0.364	1.659×	8.611×
	DAC'16	0.549	1.800	0.534	0.305	1.121×	1.028	1.272×	0.313	1.426×	5.896×
	ICCAD'21	0.728	2.100	0.745	0.347	1.275×	0.977	1.209×	0.339	1.542×	1.501×
	Ours	0.728	2.300	0.576	0.316	1.164×	**1.263**	**1.563**×	**0.400**	**1.820**×	**1.000**
Small SonicBOOM	Human Efforts	0.784	20.300	1.505	0.039	–	0.521	–	0.020	–	–
	ISCA'14	0.820	15.000	1.284	0.055	1.415×	0.638	1.226×	0.035	1.735×	5.803×
	DAC'16	0.808	14.700	1.251	0.055	1.423×	0.645	1.239×	0.035	1.764×	4.792×
	ICCAD'21	0.847	20.000	1.503	0.042	1.097×	0.564	1.082×	0.024	1.187×	1.305×
	Ours	0.840	15.200	1.254	**0.055**	**1.432**×	**0.670**	**1.287**×	**0.037**	**1.843**×	**1.000**
Medium SonicBOOM	Human Efforts	1.194	25.600	1.933	0.047	–	0.618	–	0.029	–	–
	ISCA'14	1.236	19.600	1.624	**0.063**	**1.353**×	**0.761**	**1.233**×	**0.048**	**1.667**×	5.688×
	DAC'16	1.376	25.400	1.925	0.054	1.161×	0.715	1.157×	0.039	1.344×	4.697×
	ICCAD'21	1.445	27.100	2.158	0.053	1.144×	0.670	1.084×	0.036	1.240×	1.279×
	Ours	1.287	20.600	1.735	0.062	1.340×	0.742	1.201×	0.046	1.610×	**1.000**
Large SonicBOOM	Human Efforts	1.487	44.600	3.206	0.033	–	0.464	–	0.015	–	–
	ISCA'14	1.490	30.900	2.542	0.048	1.446×	0.586	1.263×	0.028	1.827×	5.892×
	DAC'16	1.492	32.400	2.674	0.046	1.381×	0.558	1.202×	0.026	1.661×	4.865×
	ICCAD'21	1.916	40.900	3.672	0.047	1.405×	0.522	1.125×	0.024	1.581×	1.325×
	Ours	1.588	31.400	2.564	**0.051**	**1.517**×	**0.619**	**1.335**×	**0.031**	**2.025**×	**1.000**
Mega SonicBOOM	Human Efforts	1.950	57.800	4.806	0.034	–	0.406	–	0.014	–	–
	ISCA'14	2.496	56.600	5.368	0.044	1.307×	0.465	1.146×	0.021	1.498×	5.544×
	DAC'16	2.500	56.200	5.380	0.044	1.318×	0.465	1.145×	0.021	1.510×	4.578×
	ICCAD'21	2.482	60.700	4.701	0.041	1.212×	**0.528**	**1.301**×	0.022	1.578×	1.247×
	Ours	2.523	55.700	5.251	**0.045**	**1.343**×	0.480	1.184×	**0.022**	**1.590**×	**1.000**
Giga SonicBOOM	Human Efforts	1.872	71.600	5.069	0.026	–	0.369	–	0.010	–	–
	ISCA'14	2.253	62.200	6.001	0.036	1.386×	0.375	1.017×	0.014	1.409×	5.632×
	DAC'16	2.252	77.300	5.600	0.029	1.115×	**0.402**	**1.089**×	0.012	1.214×	4.651×
	ICCAD'21	2.265	74.500	5.865	0.030	1.163×	0.386	1.046×	0.012	1.216×	1.267×
	Ours	2.269	59.500	5.746	**0.038**	**1.459**×	0.395	1.070×	**0.015**	**1.560**×	**1.000**

[1] "–" denotes not applicable.

Table 2: Comparison with Human Efforts and Prior Arts

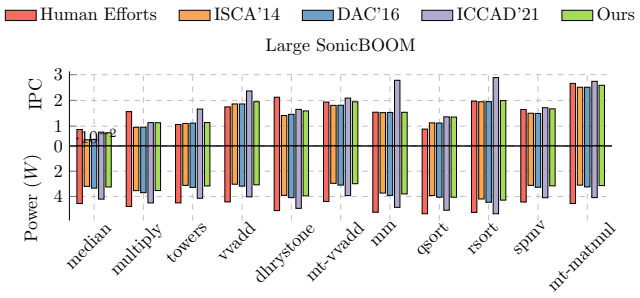

Figure 9: Analysis w. more workloads.

find those solutions using an average of 4130.89 seconds, i.e., 4.07× higher efficiency than baselines. For Medium SonicBOOM, power/area efficiency is comparable since our solution trades 4.13% more performance than ISCA'14.

Analysis w. More Workloads

We analyze explored microarchitectures with more workloads to study how RL solutions outperform other methods and human implementations. Figure 9 lists related results for Large SonicBOOM. The areas for human implementations, ISCA'14, DAC'16, ICCAD'21 and ours are 3.21, 2.54, 2.67, 3.67, and 2.56 in mm^2, respectively. While human implementations and the ICCAD'21 solution achieve higher performance on most workloads, they require more area and dissipate more power, resulting in a lower performance per watt or PPA trade-off. Compared to human implementations, our solution demonstrates significant improvements in three metrics by factors of 1.35×, 1.23×, and 1.66×, respectively. Additionally, it outperforms the baselines with average improvements of 7.74%, 12.47%, and 21.15% in the corresponding metric values. Our solution adopts a Gshare branch predictor, 16KB I-cache, and 32KB D-cache. The PPA trade-off is further enhanced by increasing integer issue queue sizes and removing redundant resources. Our solution achieves a superior PPA trade-off by selecting a more suitable branch predictor, cache structures, and a balanced allocation of resources for queues, stacks, and buffers.

Conclusion

We propose an RL solution to deal with automated RISC-V microarchitecture design. Our solution removes unrealistic assumptions and is tightly coupled with expert knowledge. Experiments show that our method on RISC-V Rocket and SonicBOOM achieves an average of 16.03% PPA trade-off improvement over prior state-of-the-art approaches with 4.07× higher efficiency.

Acknowledgements

This work is supported in part by National Key R&D Program of China (2022YFB2901100) and The Research Grants Council of Hong Kong SAR (No. CUHK14210723).

References

Abels, A.; Roijers, D.; Lenaerts, T.; Nowé, A.; and Steckelmacher, D. 2019. Dynamic Weights In Multi-objective Deep Reinforcement Learning. In *International Conference on Machine Learning (ICML)*, 11–20. PMLR.

Asanović, K.; Avizienis, R.; Bachrach, J.; et al. 2016. The Rocket Chip Generator. Technical report, University of California, Berkeley.

Bai, C.; Huang, J.; Wei, X.; Ma, Y.; Li, S.; Zheng, H.; Yu, B.; and Xie, Y. 2023a. ArchExplorer: Microarchitecture Exploration Via Bottleneck Analysis. In *IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 268–282.

Bai, C.; Sun, Q.; Zhai, J.; Ma, Y.; Yu, B.; and Wong, M. 2021. BOOM-Explorer: RISC-V BOOM Microarchitecture Design Space Exploration Framework. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 1–9.

Bai, C.; Sun, Q.; Zhai, J.; Ma, Y.; Yu, B.; and Wong, M. D. 2023b. BOOM-Explorer: RISC-V BOOM Microarchitecture Design Space Exploration. *ACM Transactions on Design Automation of Electronic Systems (TODAES)*.

Binkert, N.; Beckmann, B.; Black, G.; et al. 2011. The Gem5 Simulator. *SIGARCH Comput. Archit. News*, 39(2): 1–7.

Carlson, T. E.; Heirman, W.; and Eeckhout, L. 2011. Sniper: Exploring the level of abstraction for scalable and accurate parallel multi-core simulation. In *ACM/IEEE Supercomputing Conference (SC)*, 1–12.

Carlson, T. E.; Heirman, W.; Eyerman, S.; Hur, I.; and Eeckhout, L. 2014. An Evaluation of High-level Mechanistic Core Models. *ACM Transactions on Architecture and Code Optimization (TACO)*, 11(3): 1–25.

Chen, C.; Xiang, X.; Liu, C.; Shang, Y.; Guo, R.; Liu, D.; Lu, Y.; Hao, Z.; Luo, J.; Chen, Z.; et al. 2020. Xuantie-910: A commercial multi-core 12-stage pipeline out-of-order 64-bit high performance risc-v processor with vector extension: Industrial product. In *IEEE/ACM International Symposium on Computer Architecture (ISCA)*, 52–64.

Chen, T.; and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 785–794.

Chen, T.; Guo, Q.; Tang, K.; Temam, O.; Xu, Z.; Zhou, Z.-H.; and Chen, Y. 2014. Archranker: A ranking approach to design space exploration. In *IEEE/ACM International Symposium on Computer Architecture (ISCA)*, 1–12.

Clark, L. T.; Vashishtha, V.; Shifren, L.; Gujja, A.; Sinha, S.; Cline, B.; Ramamurthy, C.; and Yeric, G. 2016. ASAP7: A 7-nm FinFET Predictive Process Design Kit. *Microelectronics Journal*, 53: 105–115.

Eyerman, S.; Eeckhout, L.; Karkhanis, T.; and Smith, J. E. 2009. A Mechanistic Performance Model for Superscalar Out-of-order Processors. *Transactions on Computer Systems*, 27(2): 1–37.

Grayson, B.; Rupley, J.; Zuraski, G. Z.; Quinnell, E.; Jiménez, D. A.; Nakra, T.; Kitchin, P.; et al. 2020. Evolution of the Samsung Exynos CPU Microarchitecture. In *IEEE/ACM International Symposium on Computer Architecture (ISCA)*, 40–51. IEEE.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft Actor-Critic: Off-policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *International Conference on Machine Learning (ICML)*, 1861–1870. PMLR.

Karkhanis, T. S.; and Smith, J. E. 2007. Automated design of application specific superscalar processors: an analytical approach. In *IEEE/ACM International Symposium on Computer Architecture (ISCA)*, 402–411.

Lee, Y.; Waterman, A.; Cook, H.; Zimmer, B.; Keller, B.; Puggelli, A.; Kwak, J.; Jevtic, R.; Bailey, S.; Blagojevic, M.; et al. 2016. An Agile Approach to Building RISC-V Microprocessors. *IEEE Micro*, 36(2): 8–20.

Li, D.; Yao, S.; Liu, Y.-H.; Wang, S.; and Sun, X.-H. 2016. Efficient design space exploration via statistical sampling and AdaBoost learning. In *ACM/IEEE Design Automation Conference (DAC)*, 1–6.

Li, S.; Ahn, J. H.; Strong, R. D.; Brockman, J. B.; Tullsen, D. M.; and Jouppi, N. P. 2009. McPAT: An integrated power, area, and timing modeling framework for multicore and manycore architectures. In *IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 469–480.

Li, S.; Bai, C.; Wei, X.; Shi, B.; Chen, Y.-K.; and Xie, Y. 2022. 2022 ICCAD CAD Contest Problem C: Microarchitecture Design Space Exploration. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 1–7.

Mnih, V.; Badia, A. P.; et al. 2016. Asynchronous Methods for Deep Reinforcement Learning. In *International Conference on Machine Learning (ICML)*, volume 48, 1928–1937.

Peng, J.; and Williams, R. J. 1994. Incremental Multi-step Q-learning. In *Machine Learning Proceedings 1994*, 226–232. Elsevier.

Roijers, D. M.; and Whiteson, S. 2017. Multi-objective Decision Making. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 11(1): 1–129.

Roijers, D. M.; Whiteson, S.; and Oliehoek, F. A. 2015. Computing Convex Coverage Sets for Faster Multi-objective Coordination. *Journal of Artificial Intelligence Research*, 52: 399–443.

Schrittwieser, J.; Antonoglou, I.; Hubert, T.; Simonyan, K.; Sifre, L.; Schmitt, S.; Guez, A.; Lockhart, E.; Hassabis, D.; Graepel, T.; et al. 2020. Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model. *Nature*, 588(7839): 604–609.

Schulman, J.; Moritz, P.; Levine, S.; Jordan, M.; and Abbeel, P. 2016. High-Dimensional Continuous Control Using Generalized Advantage Estimation. In Bengio, Y.; and LeCun, Y., eds., *International Conference on Learning Representations (ICLR)*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*.

Seznec, A.; and Michaud, P. 2006. A case for (partially) TAgged GEometric history length branch prediction. *The Journal of Instruction-Level Parallelism*, 8: 23.

Smith, J. E.; and Pleszkun, A. R. 1988. Implementing Precise Interrupts in Pipelined Processors. *IEEE Transactions on Computers*, 37(5): 562–573.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing Data Using t-SNE. *Journal of Machine Learning Research (JMLR)*, 9(11).

Williams, C. K.; and Rasmussen, C. E. 2006. *Gaussian Processes for Machine Learning*, volume 2. MIT press Cambridge, MA.

Xu, Y.; Yu, Z.; Tang, D.; Chen, G.; Chen, L.; Gou, L.; Jin, Y.; Li, Q.; Li, X.; Li, Z.; et al. 2022. Towards Developing High Performance RISC-V Processors Using Agile Methodology. In *IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 1178–1199. IEEE.

Yang, R.; Sun, X.; and Narasimhan, K. 2019. A Generalized Algorithm for Multi-Objective Reinforcement Learning and Policy Adaptation. In *Annual Conference on Neural Information Processing Systems (NIPS)*.

Yu, Z.; Bai, C.; Hu, S.; Chen, R.; He, T.; Yuan, M.; Yu, B.; and Wong, M. 2023. IT-DSE: Invariance Risk Minimized Transfer Microarchitecture Design Space Exploration. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 1–9. IEEE.

Zhai, J.; Bai, C.; Zhu, B.; Cai, Y.; Zhou, Q.; and Yu, B. 2021. McPAT-Calib: A Microarchitecture Power Modeling Framework for Modern CPUs. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 1–9. IEEE.

Zhao, J.; Korpan, B.; Gonzalez, A.; and Asanovic, K. 2020. SonicBOOM: The 3rd Generation Berkeley Out-of-order Machine. In *Workshop on Computer Architecture Research with RISC-V (CARRV)*.